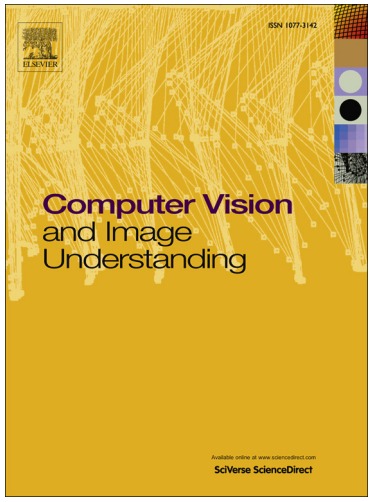


SHOT: Unique Signatures of Histograms for Surface and Texture Description

Samuele Salti, Federico Tombari, Luigi Di Stefano

PII: S1077-3142(14)00098-8
DOI: http://dx.doi.org/10.1016/j.cviu.2014.04.011
Reference: YCVIU 2133

To appear in: *Computer Vision and Image Understanding*

Received Date: 31 January 2013
Accepted Date: 27 April 2014

Please cite this article as: S. Salti, F. Tombari, L.D. Stefano, SHOT: Unique Signatures of Histograms for Surface and Texture Description, *Computer Vision and Image Understanding* (2014), doi: http://dx.doi.org/10.1016/j.cviu.2014.04.011

SHOT: Unique Signatures of Histograms for Surface and Texture Description

Samuele Salti, Federico Tombari, Luigi Di Stefano

Department of Computer Science and Engineering, University of Bologna, Italy

Abstract

This paper presents a local 3D descriptor for surface matching dubbed SHOT. Our proposal stems from a taxonomy of existing methods which highlights two major approaches, referred to as *Signatures* and *Histograms*, inherently emphasizing descriptiveness and robustness respectively. We formulate a comprehensive proposal which encompasses a repeatable local reference frame as well as a 3D descriptor, the latter featuring an hybrid structure between *Signatures* and *Histograms* so as to aim at a more favorable balance between descriptive power and robustness. A quite peculiar trait of our method concerns seamless integration of multiple cues within the descriptor to improve distinctiveness, which is particularly relevant nowadays due to the increasing availability of affordable RGB-D sensors which can gather both depth and color information. A thorough experimental evaluation based on datasets acquired with different types of sensors, including a novel RGB-D dataset, vouches that SHOT outperforms state-of-the-art local descriptors in experiments addressing descriptor matching for object recognition, 3D reconstruction and shape retrieval.

Keywords: Surface Matching, 3D Descriptors, Object Recognition, 3D Reconstruction

1. Introduction

Automatic recognition of shapes in 3D data, also referred to as *surface matching*, attracts growing interest in the research community, with application to areas such as shape retrieval, 3D reconstruction, object recognition/categorization, manipulation and grasping, robot localization and navigation. A key enabling factor for the development of this technology is

represented by the availability of low-cost 3D sensors. Moreover, the majority of such sensors can acquire not only the 3D shape of the scene, but also its texture (the so-called RGB-D data): this is the case, *e.g.*, of stereo sensors, structure-from-motion systems, certain laser scanners as well as the recently introduced Microsoft *Kinect* and Asus *Xtion* devices.

Surface matching is usually tackled either by means of a global or a local approach. According to the former, a feature describes the whole surface, whereas the latter relies on local keypoints and regional feature descriptions to determine point-to-point correspondences between surfaces. Borrowing a denomination typical of the face recognition community [1], we refer here to these two approaches as, respectively, *holistic* and *feature-based*. While the holistic approach is popular in the context of 3D *shape retrieval* (see for example [2, 3, 4]), feature-based methods are inherently more effective for 3D *object recognition* due to the ability to withstand clutter and occlusions.

Feature-based methods may rely on 3D keypoints extracted from surfaces. This task is accomplished by 3D detectors [5, 6, 7], whose aim is to single out points that are distinctive, to allow for effective description and matching, as well as repeatable with respect to vantage point variations and noise. A characteristic scale may also be associated to each keypoint, so as to provide the following description stage with a local neighborhood size [5, 8, 9, 10, 11]. A performance evaluation of 3D keypoint detection algorithms has been recently proposed in [12]. Information within the local neighborhood of each keypoint is encoded by means of a 3D descriptor, so as to obtain a compact local representation of the input data invariant up to a predefined transformation (rotation, translation, scaling, point density variations, ...). The majority of descriptors relies on a repeatable local Reference Frame (RF), whose importance for the overall descriptor performance has been demonstrated in the preliminary version of this work [13] as well as in another recent work, entirely devoted to the study of local RFs [14]. Descriptors are then matched across different views to attain point-to-point correspondences.

In this paper we first propose a categorization of the main proposals for 3D description (Sec. 2), by dividing state-of-the-art methods between *Signatures* and *Histograms*. We then present our proposal for a robust local RF that is unique and repeatable (Sec. 3). Starting from the analysis of state-of-the-art descriptors provided in our taxonomy, in Sec. 4 we propose the Signature of Histograms of OrienTations (SHOT) descriptor as an attempt to leverage on the benefits of both Signatures and Histograms. We then present

the generalization of SHOT to the case of *RGB-D* data, by showing how its formulation can seamlessly and efficiently incorporate the description of color channels. Finally, we provide a thorough experimental evaluation of our proposal, addressing tasks such as descriptor matching in presence of noise, clutter, occlusion and point density variations, automatic 3D reconstruction from multiple unordered views and 3D shape retrieval. Throughout the experiments, we compare SHOT to several state-of-art methods and consider datasets acquired with different sensors. As part of our evaluation methodology, we also propose a new *RGB-D* dataset with ground-truth comprising a large number of scene and model views acquired by a Kinect device.

This paper extends and consolidates the research work previously presented in [13] and [15]: the two algorithms are now presented together as a unique algorithm and all the details and pseudo-code to reproduce the method are provided; we enrich and revise the taxonomy of 3D descriptor and, for the sake of clarity and generalization, drop the distinction between unique and unambiguous local RFs as well as highlight the ability of each proposal to handle RGB-D data; we extend the comparative evaluation concerning description matching experiments by including three additional prominent proposals, *i.e.* FPFH [16], MeshHoG [11] and KP [5]; descriptor matching experiments are carried out also by comparing all methods using the same algorithm to compute the local RF, so as to better highlight the contribution given by both the proposed local RF as well as the proposed descriptor; we propose a novel RGB-D dataset with ground-truth, acquired with the Kinect sensor, on which we compare our descriptor to other shape and color descriptors; we demonstrate full-3D reconstruction from unordered views based on Kinect data, which are noisier than those used in [13]; we also address a Shape Retrieval scenario to evaluate and compare quantitatively our proposal with respect to other descriptors in another relevant application.

2. Taxonomy of 3D descriptors

In Table 1 we propose a categorization of the main approaches for 3D description. As shown in the second column, we divide 3D descriptors into two main categories, namely *Signatures* and *Histograms* (Figure 1). The methods in the first category, which mostly includes earliest works on this subject, describe the 3D surface neighborhood of a given point (hereinafter *support*) by defining an invariant local Reference Frame and encoding, according to the local coordinates, one or more geometric measurements computed indi-

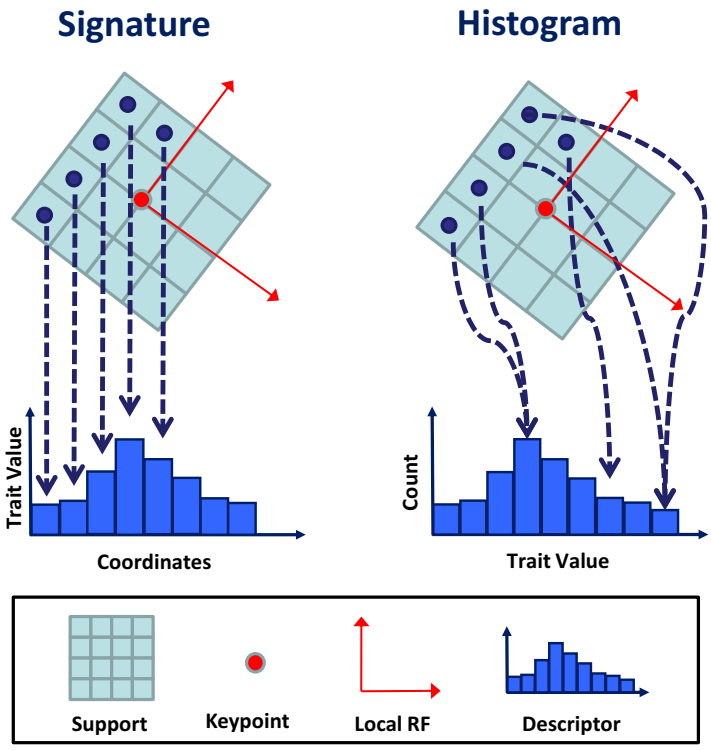


Figure 1: The two main classes of 3D descriptors: signatures and histograms. For the sake of clarity, the figure reports an example in a 2D domain, wherein the categorization is still valid.

vidually on each point of a subset of the support. Signatures are potentially highly descriptive thanks to the use of spatially well-localized information, but small errors in the definition of the local RF or small perturbations in the encoded trait can substantially modify the final descriptor. On the other hand, Histograms describe the support by encoding counters of local topological entities (*e.g.* vertices, mesh triangle areas) into histograms according to a specific quantized domain (*e.g.* point coordinates, curvatures, normal angles). If the descriptor domain is based on coordinates, then also histogram-based methods require the definition of a local RF (*e.g.* 3D Shape Context [17], Tensor [18]). For other domains, *e.g.* angles between normals as in LSP or cylindric coordinates as in Spin Images [19], only a repeatable Reference Axis (RA) is required. In broad terms, when compared to Signa-

tures, Histograms trade-off descriptive power for robustness by compressing information into bins.

As far as Signatures are concerned, one of the first proposals is *Structural Indexing* [20], which builds up a representation based on either a *3D curve* or a *Splash* depending on the characteristics of the 3D support. The former encodes the angles between consecutive segments of the polygonal approximation of edges (corresponding to depth or orientation discontinuities) on the surface. The latter encodes as a 3D curve the local distribution of surface orientations along a geodesic circle centered on the point. In *Point Signatures* [21] the signature is given by the signed height of the 3D curve obtained by intersecting a sphere centered at the point with the surface. *3D Point Fingerprint* [22] encodes normal angle variations and contour radius variations along different geodesic circles projected onto the tangent plane. The work in [8] proposes a descriptor that encodes the components of the normals within the support by deploying a 2D parametrization of the local surface based on the exponential mapping. In [5], the signature is given by the third coordinate of each vertex of the support expressed in the local RF. This scheme has been recently extended to a hybrid scheme, by computing a SIFT descriptor out of the resulting depth image [23]. Finally, [24] extends the successful SURF descriptor [25] to 3D data, by voxelizing the 3D mesh so as to efficiently compute Haar wavelets and use their responses as signature trait.

As for Histograms, those relying on the definition of just a RA are typically based on the feature point normal. For example, *Spin Images* [19] computes 2D histograms of points falling within a cylindrical volume by means of a plane that "spins" around the normal. Within the same subclass, *Local Surface Patches* [6] computes histograms of normals and *shape indexes* [26] of the points belonging to the support. As for methods relying on the definition of a full local RF, *3D Shape Context* [17] modifies the basic idea of Spin Images by accumulating 3D histograms of points within a sphere centered at the feature point. *Intrinsic Shape Signatures* [7] proposes an improvement of 3D Shape Context based on a different partitioning of the 3D local volume as well as on a different definition of the local RF. *Point Feature Histograms* (PFH) [27] and *Fast Point Feature Histograms* (FPFH) [16] accumulate in a 3D histogram three angular values computed between pairs of points falling within the support and their respective normals. Finally, [18] accumulate 3D histograms (*Tensors*) of mesh triangle areas within a cubic support.

Although not explicitly grounded on our taxonomy, the recent MeshHoG

(MH) 3D descriptor [11] shares the same hybrid structure as SHOT. It encodes several histograms of gradients of a scalar function defined at each vertex (*e.g.* the mean curvature) ordered according to a local RF. Moreover, likewise SHOT and differently from all the other methods reviewed here, MeshHog allows for inclusion of both color as well as 3D shape information within the descriptor. Another hybrid descriptor is the recent Rotational Projection Statistics (RoPS) [28], which encodes robust statistics (central moments and entropy) of the distributions of vertices of the mesh, rotated along the RF axes, and projected and quantized on the 2D xy, yz, and zx planes.

A few descriptors, though, can be interpreted neither as Signatures nor as Histograms. Among them, the most relevant proposal is the Heat Kernel Signature [29]. The descriptor is obtained by a restriction of the heat kernel to the time dimension. The heat kernel fully characterizes the underling manifold up to isometries. The descriptor is inherently multi-scale and does not make use of a local Reference Frame.

3. Local RF from disambiguated EVD

The definition of a local RF, invariant to translations and rotations and robust to noise and clutter, has been the preferred option to endow a 3D descriptor with invariance to the same sources of variations, similarly to the way rotation and/or scale invariance is injected into 2D descriptors. On the other hand, the definition of such a canonical frame is challenging. As highlighted in the third column of Table 1, none of current local RF proposed together with descriptors is unique, except that of MeshHoG and 3D SURF. However, the local RF defined within the MeshHoG descriptor is highly sensitive to noise, as shown in [14]. As pointed out in [30], also the local RF of 3D SURF is not robust, so that its deployment in a voting scheme *á la* ISM [31] requires casting votes for multiple locations in order to counteract its low repeatability. Resorting to multiple local RFs, hence multiple descriptions of the given neighborhood, is suboptimal as it implies a higher cost associated with the description process in terms of both execution time and

[1]The original algorithm in [8] does not provide a disambiguation for the sign of the proposed local RF, while the publicly available implementation does include an explicit disambiguation step.

Table 1: Taxonomy of 3D descriptors.

Method	Category	Unique LRF	Color
StInd [20]	Sign.	No	No
PS [21]	Sign.	No	No
3DPF [22]	Sign.	No	No
3DGSS [8]	Sign.	No[1]	No
KP [5]	Sign.	No	No
LD-SIFT [23]	Sign.	No	No
3D-SURF [24]	Sign.	Yes	No
SI [19]	Hist.	RA	No
LSP [6]	Hist.	RA	No
3DSC [17]	Hist.	No	No
ISS [7]	Hist.	No	No
PFH [27]	Hist.	RA	No
FPFH [16]	Hist.	RA	No
Tensor [18]	Hist.	No	No
HKS [29]	Other	-	No
RoPS [28]	Both	Yes	No
MH [11]	Both	Yes	Yes
SHOT	**Both**	**Yes**	**Yes**

memory occupancy. Moreover, the matching stage becomes more ambiguous and slower.

We have designed and extensively tested a variety of novel local RFs in order to get to a unique and robust proposal. We present here the method that turned out the most robust in our experimental evaluation. Unlike a recent study on local RFs addressing partial views registration [14], the design and evaluation of our algorithm has been tailored mainly to surface matching in presence of clutter and occlusions. It builds on a well known technique presented in [32] and [33], where the problem of normal estimation in presence of noise is specifically addressed. A Total Least Squares (TLS) estimation of the normal direction is obtained in Hoppe's and Mitra's works [32], [33] by EigenValue Decomposition (EVD) of the covariance matrix, $\mathbf{M}$, of the $k-$nearest neighbors p_i of the point:

$$\mathbf{M} = \frac{1}{k}\sum_{i=0}^{k}(\mathbf{p}_i - \hat{\mathbf{p}})(\mathbf{p}_i - \hat{\mathbf{p}})^T,\ \hat{\mathbf{p}} = \frac{1}{k}\sum_{i=1}^{k}\mathbf{p}_i\ . \tag{1}$$

In particular, the TLS estimation of the normal direction is given by the eigenvector corresponding to the smallest eigenvalue of M. Finally, they

perform the sign disambiguation of the normals *globally* by means of sign consistency, *i.e.* by propagating the sign from a seed chosen heuristically. While this has proven to be a robust and effective technique for surface reconstruction of a single object, it can not be as effective for local surface description since in the latter case signs must be repeatable across any possible object pose as well as in scenes with multiple objects, so that a *local* rather than global sign disambiguation method is mandatory. Moreover, Hoppe's sign disambiguation concerns the normal only, hence it leaves ambiguous the signs of the remaining two axes.

In our proposal, we modify (1) to assign distant points smaller weights, so as to increase repeatability in presence of clutter. Then, to improve robustness to noise, all points laying within the spherical support (of radius R) which are used to compute the descriptor are also used to calculate $\mathbf{M}$. For the sake of efficiency, we neglect the centroid computation, replacing it with the feature point $\mathbf{p}$. Therefore, we calculate M as a weighted linear combination,

$$\mathbf{M} = \frac{1}{\sum\limits_{i:d_i \leq R} (R - d_i)} \sum_{i:d_i \leq R} (R - d_i)(\mathbf{p}_i - \mathbf{p})(\mathbf{p}_i - \mathbf{p})^T \tag{2}$$

where $d_i = \|\mathbf{p}_i - \mathbf{p}\|_2$. Our experimental evaluation indicates that the eigenvectors of $\mathbf{M}$ define repeatable, orthogonal directions in presence of noise and clutter. It is worth pointing out that, compared to [32] and [33], in our proposal the third eigenvector no longer represents the TLS estimation of the normal direction and sometimes it notably differs from it. However, this does not affect performance, as in local surface description the definition of a highly repeatable and robust triplet of orthogonal directions is more important than its geometrical or topological meaning.

Eigenvectors of (2) provides repeatable directions for the local RF axes, but they need to be disambiguated to yield a unique local RF. The problem of sign disambiguation for EVD and SVD has been recently addressed in [34]. Their proposal basically reorients each singular vector or eigenvector so that its sign is coherent with the majority of the vectors it is representing. We determine the sign on the local $\mathbf{x}$ and $\mathbf{z}$ axes according to this principle. In the following we refer to the three unit eigenvectors in decreasing eigenvalue order as the $\mathbf{x}^+$, $\mathbf{y}^+$ and $\mathbf{z}^+$ axis, respectively. With $\mathbf{x}^-$, $\mathbf{y}^-$ and $\mathbf{z}^-$, we denote instead the opposite unit vectors. Let $M(k)$ be the subset of points

within the support whose distance from the feature point is among the k closest to the median distance d_m, *i.e.*

$$M(k) \doteq \{i : |m - i| \leq k, m = \arg\operatorname*{median}_j d_j\} \ . \tag{3}$$

Then, the final disambiguated $\mathbf{x}$ axis is defined as

$$S_x^+ \doteq \left\{i : d_i \leq R \ \wedge \ (\mathbf{p}_i - \mathbf{p}) \cdot \mathbf{x}^+ \geq 0\right\} \tag{4}$$

$$S_x^- \doteq \left\{i : d_i \leq R \ \wedge \ (\mathbf{p}_i - \mathbf{p}) \cdot \mathbf{x}^- > 0\right\} \tag{5}$$

$$\tilde{S}_x^+ \doteq \left\{i : i \in M(k) \ \wedge \ (\mathbf{p}_i - \mathbf{p}) \cdot \mathbf{x}^+ \geq 0\right\} \tag{6}$$

$$\tilde{S}_x^- \doteq \left\{i : i \in M(k) \ \wedge \ (\mathbf{p}_i - \mathbf{p}) \cdot \mathbf{x}^- > 0\right\} \tag{7}$$

$$\mathbf{x} = \begin{cases} \mathbf{x}^+, & |S_x^+| > |S_x^-| \\ \mathbf{x}^-, & |S_x^+| < |S_x^-| \\ \mathbf{x}^+, & |S_x^+| = |S_x^-| \ \wedge \ |\tilde{S}_x^+| > |\tilde{S}_x^-| \\ \mathbf{x}^-, & |S_x^+| = |S_x^-| \ \wedge \ |\tilde{S}_x^+| < |\tilde{S}_x^-| \end{cases} \tag{8}$$

To disambiguate the EVD also at those points where $|S_x^+| = |S_x^-|$, we consider only an odd number k of vertices in $M(k)$, yielding the subsets $\tilde{S}_x^+$ and $\tilde{S}_x^-$, and we reorient the eigenvector so that its sign is coherent with the majority of such vectors. The same procedure is used to disambiguate the $\mathbf{z}$ axis. Finally, the $\mathbf{y}$ axis is obtained as $\mathbf{z} \times \mathbf{x}$.

4. SHOT descriptor

In Sec. 2, we classified 3D descriptors as either histograms or signatures. We have designed our proposal following this intuition and aiming at a local representation that is efficient, descriptive, robust to noise and clutter as well as to point density variation. The point density issue is peculiar to the 3D scenario, wherein a given real world 3D volume may be represented with different amounts of vertices in its mesh approximation, *e.g.* due to the use of different 3D sensors (stereo, ToF cameras, LIDARs, etc...) and/or different sensing distances.

Beside our taxonomy, another source of inspiration has been the related field of 2D feature descriptors, which has reached a remarkable maturity during the last years. By analyzing SIFT [35], arguably the most successful and widespread proposal among 2D descriptors, we have singled out the

major reasons behind its effectiveness. Histograms are used in several steps of the algorithm, from the definition of the local orientation to the descriptor itself, and contribute to the robustness of the method. As a single histogram computed on the whole patch would be not descriptive enough, SIFT relies on a set of local histograms that are computed on specific subsets of pixels defined by a regular grid superimposed on the patch. The use of this coarse geometric grid creates a signature-like structure. Finally, the local histograms encode first order derivatives of the signal, *i.e.* intensity gradients.

Based on these considerations, we propose a 3D descriptor that encodes histograms of the normals within the support (*i.e.* first-order differential entities). The discriminative power of the descriptor is enhanced by introducing geometric information concerning the location of the points within the support, thereby mimicking a signature. This is done by computing a set of local histograms over the 3D volumes defined by a 3D grid superimposed on the support. The grid is aligned with the axes defined by the local RF introduced in the previous section. Since our descriptor lays at the intersection between Histograms and Signatures, we dub it Signature of Histograms of OrienTations (SHOT).

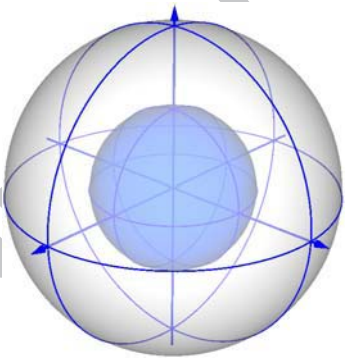
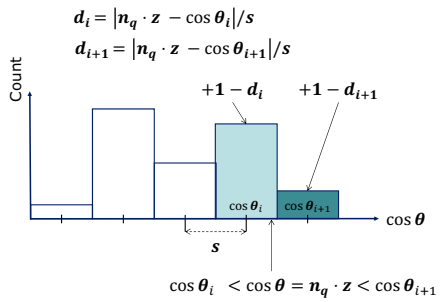


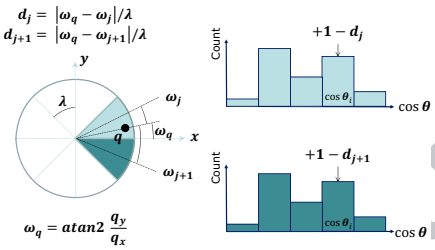
Figure 2: Signature structure for SHOT.

For each of the local histograms, we accumulate a point into bins according to the cosine of the angle θ_q between the normal at the point, $\mathbf{n}_q$, and the local $\mathbf{z}$ axis at the feature point, $\mathbf{z}_k$. The reason to use the cosine is twofold: it can be computed rapidly, since $\cos\theta_q = \mathbf{z}_k \cdot \mathbf{n}_q$; an equally spaced binning on $\cos\theta_q$ is equivalent to a spatially varying binning on θ_q, whereby a coarser binning is created for directions close to the reference normal direction and a finer one for orthogonal directions. In this way, small differences in orthogonal directions to the normal, *i.e.* presumably the most informative ones, cause points to be accumulated in different bins leading to different histograms. Moreover, in presence of quasi-planar regions (*i.e.* not very descriptive ones) this choice limits histogram differences due to noise by concentrating counts into a fewer number of bins.

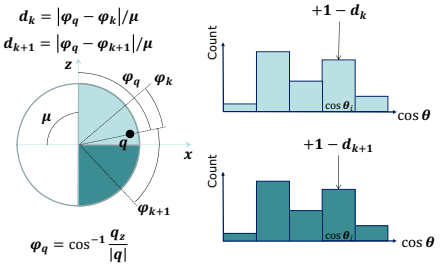
As for the signature structure, we use an isotropic spherical grid that encompasses partitions along the radial, azimuth and elevation axes, as sketched in Fig. 2. Since each volume of the grid encodes a very descriptive



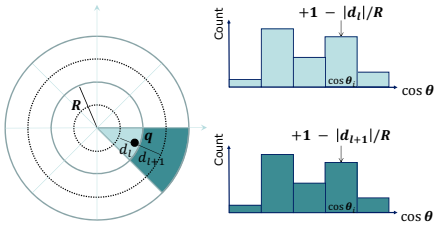
(a) Interpolation on normal cosines



(b) Interpolation on azimuth



(c) Interpolation on elevation



(d) Interpolation on distance

Figure 3: Quadrilinear interpolation to accumulate weights into histograms.

entity represented by the local histogram, we can use a coarse partitioning of the spatial grid and hence a small cardinality of the descriptor. In particular, our experimentations indicate that 32 is a proper number of volumes, resulting from 8 azimuth divisions, 2 elevation divisions and 2 radial divisions (though, for better clarity, only 4 azimuth divisions are shown in Fig. 2). Combined with the fact that our tuning experiments indicate a proper number of bins for the internal histograms to be 11, we obtain a total descriptor length of 352, a value that allows for faster indexing and matching with respect to the length of 3DSC (1980) or ISS (595). We are aware of an allegedly more effective, proposal for sphere subdivision presented in [7]. However, the author ascribes the higher effectiveness of her method to the absence of degenerating points in the division, which instead are present at the poles and center of the subdivision in Fig. 2. Since our proposal requires much less bins and, as explained below, we use interpolation among bins, we found degenerating points to represent a minor issue.

Since our descriptor is based upon local histograms, it is important to avoid boundary effects, as pointed out *e.g.* in [19] and [35]. Furthermore, due to the signature structure, boundary effects may also occur due to perturbations of the local RF. Therefore, for each point being accumulated into a specific local histogram bin, we perform quadrilinear interpolation with its neighbors, i.e. the neighboring bin in the local histogram and the bins having the same index in the local histograms corresponding to the neighboring subdivisions of the grid. In particular, each bin is incremented by a weight of $1 - d$ for each dimension. As for the local histogram, d is the distance of the current entry from the central value of the bin. As for elevation and azimuth, d is the angular distance of the entry from the central value of the volume. Along the radial dimension, d is the Euclidean distance of the entry from the central value of the volume. Along each dimension, d is measured in units of the histogram or grid spacing, i.e. it is normalized by the distance between two neighbor bins or volumes. Figure 3 provides a graphic description of the quadrilinear interpolation process.

To achieve robustness to point density variations, we normalize the whole descriptor to have Euclidean norm equal to 1. This is preferable to the solution proposed in [17], i.e. normalizing each bin by the inverse of the local point density and bin volume. In fact, while [17] implicitly assumes that the sampling density may vary independently in every bin, and thus discards as not informative the differences in point density among bins, we assume global (or at least regional) variations of the density and keep the local differences

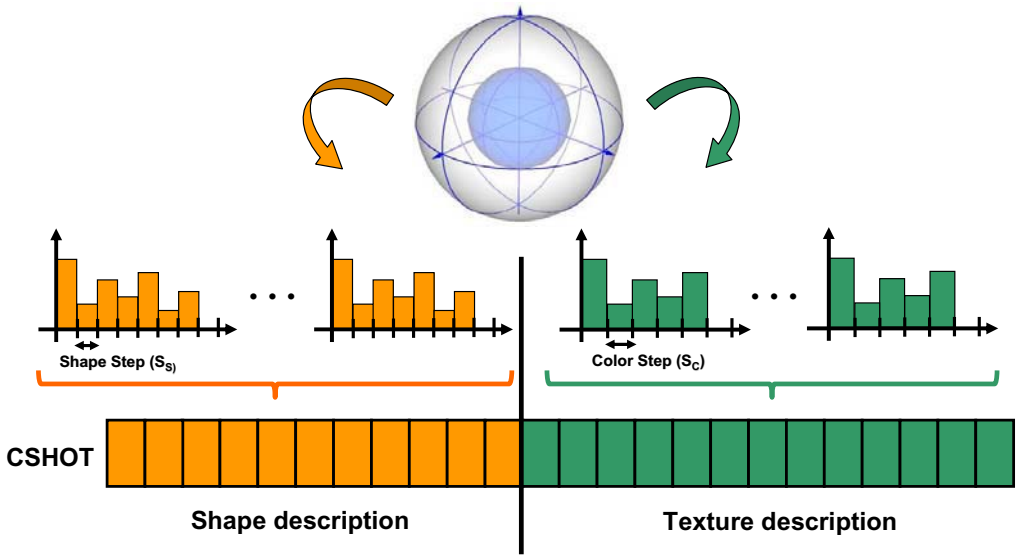


Figure 4: The proposed extension to RGB-D data merges together two signature of histograms, of shape and texture-related measurements respectively.

as a source of discriminative information.

4.1. Extension to RGB-D data

In this section we show how the design of the SHOT descriptor can be generalized seamlessly to incorporate texture information. This results in a particularly interesting approach for carrying out surface matching tasks based on the output of modern 3D sensors capable of delivering both shape and color (usually referred to as RGB-D sensors).

In the proposed generalization (Fig. 4), 2 signatures of histograms encoding, respectively, shape and texture information are computed in the spherical support and chained together in order to build the descriptor $D(k)$ at feature point k:

$$D(k) = D_{shape}(k) \cup D_{texture}(k) \tag{9}$$

As for the shape descriptor, we use the previously introduced approach, *i.e.* in every histogram we accumulate point counts according to the cosine of the angle between normals. To obtain the texture descriptor, we have to define both the point-wise trait to be compared between feature k and every

Algorithm 1 The SHOT descriptor

$\mathcal{F} = \{\text{feature points}\}$
$\mathcal{S}_{\mathbf{p}} = \{\text{points in the sphere of radius } R \text{ around } \mathbf{p}\}$
RGBData = a boolean indicating RGB-D data

compute normal $\mathbf{n_i}$ for every point $\mathbf{i}$
if RGBData **then**
  convert $\mathbf{RGB_i}$ to $\mathbf{Lab_i}$ for every point $\mathbf{i}$
end if
for $\mathbf{p} \in \mathcal{F}$ **do**
  //Computation of the local Reference Frame
  $\mathbf{M} = \mathbf{0}$
  for $\mathbf{q} \in \mathcal{S}_{\mathbf{p}}$ **do**
    $d = ||\mathbf{p} - \mathbf{q}||_2$
    $\mathbf{M} = \mathbf{M} + (R - d)(\mathbf{p} - \mathbf{q})(\mathbf{p} - \mathbf{q})^T$
  end for
  $\mathbf{M} = \mathbf{VDV}^{-1}, \mathbf{V} = [\mathbf{x}^+ \ \mathbf{y}^+ \ \mathbf{z}^+]$ *//Compute EVD*
  //Disambiguate axes
  $\mathbf{x} = \mathbf{x}^+$ **if** $|S_x^+| \geq |S_x^-|$ **else** $\mathbf{x}^-$
  $\mathbf{z} = \mathbf{z}^+$ **if** $|S_z^+| \geq |S_z^-|$ **else** $\mathbf{z}^-$
  $\mathbf{y} = \mathbf{z} \times \mathbf{x}$
  //Computation of the signature of histograms
  for $\mathbf{q} \in \mathcal{S}_{\mathbf{p}}$ **do**
    compute local coordinates (q_x, q_y, q_z) of $\mathbf{q}$ wrt $[\mathbf{x}\ \mathbf{y}\ \mathbf{z}]$
    quantize (q_x, q_y, q_z) wrt to the spatial grid
    $\theta \leftarrow \mathbf{n}_q \cdot \mathbf{z}$
    quantize θ wrt to the shape histogram bins
    if RGBData **then**
      $\gamma \leftarrow ||\mathbf{Lab}_q - \mathbf{Lab}_p||_1$
      quantize γ wrt to the texture histogram bins
    end if
    quadrilinear interpolation to accumulate $\mathbf{q}$
  end for
  normalize the descriptor to Euclidean norm 1
end for

point within the support as well as a suitable metric to compare two such texture-related values.

The most intuitive choice for the texture-based point-wise trait is the RGB triplet associated with each vertex. To properly compare RGB triplets, one option is to deploy again the approach taken to describe shape, i.e. computation of the dot product. Alternatively, we have tested a metric based on the L_p norm between two triplets. In particular, we have implemented the operator based on the L_1 norm, which consists in the sum of the absolute differences between the triplets. Moreover, we have investigated on using different color spaces rather than RGB. We have chosen the *CIELab* space given its well-known property of being more perceptually uniform than the RGB space [36]. Hence, as a different solution, the point-wise color is represented by color triplets computed in this space. Comparison between Lab triplets can be done using the same approaches as those adopted for RGB triplets, i.e. the dot product or the L_1 norm. In addition, we have also investigated on the use of more specific metrics defined for the *CIELab* color space. In particular, we have considered two metrics, known as *CIE94* and *CIE2000*, that were defined by the *CIE* Commission in 1994 and 2000 respectively (for their definitions the reader is referred to [36]). According to a thorough experimental study described in [15], the best property-metric pair turned out to be the *Lab* color space together with the L_1 norm. Hence, in the reminder of this paper we will refer to the SHOT descriptor enriched with texture information as to this particular configuration.

Given the different nature of the two signatures of histograms embedded into the extended SHOT descriptor, it is useful to allow for a different number of bins in the two histogram types. Thus, the extended descriptor includes an additional parameter, which specifies the number of bins in each texture histogram and is referred to as Color Step (S_C, see Fig. 4). As discussed for shape, it is important to deal with boundary effects and spurious descriptors differences due to point density variations when encoding texture: hence, the quadrilinear interpolation and final normalization are carried out also in the SHOT descriptor for RGB-D data. The overall algorithm for computing the SHOT descriptor both for 3D and RGB-D data is reported in Alg. 1.

5. Experimental results

This section provides a thorough experimental evaluation of SHOT by addressing 3 typical application scenarios for 3D descriptors. The first con-

cerns *descriptor matching* for object recognition, where a set of 3D models has to be matched against a dataset of scenes characterized by clutter and occlusions. Then, we experiment on *shape retrieval*, where, given a query 3D object, models belonging to its same category have to be retrieved from a 3D library. Finally, we provide results concerning *3D reconstruction*, where different 2.5D views of a given object have to be aligned together to attain the object's 3D shape. Since descriptor matching for object recognition and 3D reconstruction are increasingly carried out also on RGB-D data, in the descriptor matching and the registration experiments we consider the case of input data represented by 3D meshes and descriptors relying on shape only (*depth data*), as well as *RGB-D data*.

Table 2: Parameter values used throughout *descriptor matching* experiments. The value after the slash, where present, is used with Kinect data (Dataset 4). Radii are reported in mesh resolution units. A bold font indicates a tuned value, a plain font a default value.

	Radius (mr)	Hist. Bins	Sign. Size	Hist. Color Bins	As	Bins for RF	Gauss. Sigma	Local Max Th	Length
SHOT(S)	**15/25**	**11/6**	**32/32**	–	–	–	–	–	352/176
SHOT(S+C)	**15/30**	**11/16**	**32/32**	**30/5**	–	–	–	–	1280/640
SI	**30**	**10**	–	–	60°	–	–	–	100
PS	**10**	–	**90**	–	–	–	–	**0.75**	90
KP	**30**	–	**26 ∗ 26**	–	–	–	–	–	676
MH(S)	**25/30**	**8/8**	**12/8**	–	–	36	0.7	–	96/64
MH(S+C)	**25/25**	**8/4**	**12/4**	**28/8**	–	36	0.7	–	432/48
FPFH	**20/30**	33	–	–	–	–	–	–	33
PFHRGB	**15/20**	125	–	125	–	–	–	–	250

5.1. Descriptor matching experiments

These experiments quantitatively evaluate and compare our proposal with respect to state-of-the-art approaches on different datasets. In each experiment, given a set of models and a scene characterized by clutter and occlusions, the goal is to establish correct correspondences between the features extracted from the scene and those extracted from each model. Hence, for each experiment, the first step is represented by the extraction of features from the current model and the current scene. For a fair comparison, we use

the same feature detector with all descriptors. In particular, we randomly extract a set of feature points from each model, then we extract their corresponding points from the scene, so that performance of the descriptors is not affected by errors of the detector. Finally, features are described using the evaluated descriptors and point-to-point correspondences are established by matching each scene feature to all model features. More specifically, as proposed in [35] we compute the ratio between the nearest neighbor and the second best: if the ratio is below a threshold a correspondence is established between the scene feature and its closest model feature. According to the methodology for evaluation of 2D descriptors recommended in [37], we provide results in terms of *Recall* versus *1-Precision* curves. This choice is preferable to ROC curves (i.e. *True Positive Rate* versus *False Positive rate*) when comparing descriptors or detectors due to the ambiguity in calculating the *False Positive Rate* [38].

In addition to quantitative results related to descriptor matching, we also compare proposals in terms of efficiency, providing indications concerning their computational cost: this will be discussed in Sect. 5.1.3.

5.1.1. Experiments on depth data

SHOT is compared to five state-of-the-art approaches: *Spin Images* (SI) [19] and *Fast Point Feature Histogram* (FPFH), as representatives of Histogram-based methods due to their vast popularity in the addressed scenario; the algorithm presented in [5] (referred to here as *Keypoint Matching* (KP)) and *Point Signatures* (PS) [21] as representatives of Signature-based methods, the former chosen as it is very recent, the latter given its relevance in literature; *MeshHoG* (MH) [11] is also considered since it is another *hybrid* method described in literature. All methods were implemented in C++, except for MeshHoG, whose C++ implementation is provided by the authors, FPFH, whose C++ implementation is present in the popular open-source *Point Cloud Library* (PCL)[2], and KP, which uses a MATLAB script, *i.e. gridfit* [39], to perform point density normalization and mesh smoothing. As previously mentioned, we address here descriptor matching based on shape only and, hence, all evaluated algorithms (in particular, SHOT and MeshHoG) do not make use of RGB information when available.

We have run two different types of experiment. In the first, hereinafter

[2] `www.pointclouds.org`

referred to as *Original Local RF*, we use original proposals both to define the Repeatable Axis or local RF as well as to compute the descriptor. To single out the contribution of the actual description scheme with respect to the algorithm adopted to establish a canonical reference, in the second experiment we run the evaluation using the same local RF with all methods (*i.e. Same Local RF*): in particular, we use the local RF proposed in Sec. 3. In both experiments, we use for every algorithm the matching measure that was originally proposed by its authors.

In each test, we use the three datasets proposed in [13], which can be downloaded from the SHOT project page[3]. Dataset 1 includes 6 models from the *Stanford 3D Scanning Repository*[4], 45 scenes built by randomly rotating and translating different subsets of the model set so as to create clutter and by adding Gaussian random noise with increasing standard deviation, namely σ_1 and σ_3 at respectively 10% and 30% of the average mesh resolution (computed on all models). The scenes in this dataset contain complex shapes, rich of details, corrupted by significant amount of noise. Therefore, such data favor methods aimed at robustness more than descriptive power (descriptors tend to be distinctive even though some details are lost or compressed into bins). Dataset 2 consists of the scenes in Dataset 1 (with noise σ_1) resampled down to 1/8 of their original point density. For a fair comparison in this experiment, our implementation of SI (used throughout the evaluation) normalizes each descriptor to the unit vector to make it more robust to point density variations, as proposed in [40]. Finally, Dataset 3 consists of scenes and models acquired by means of a 3D sensing technique known as *Spacetime Stereo* [41, 42]. Fig. 5c shows two scenes belonging to this Dataset, together with the two models sought for. Compared to Stanford models, surfaces of the objects appearing in Spacetime scenes are smoother, thus harder to discriminate, and less noisy than those appearing in the scenes of Dataset 1. Hence, Dataset 3 calls for high descriptive power rather than robustness.

For each of the three datasets, 1000 feature points were extracted from each model. As for the scenes, in Datasets 1 and 2 we extract $n*1000$ features per scene (n being the number of models present in the scene), whereas in Dataset 3 we extract 3000 features per scene.

[3] `www.vision.deis.unibo.it/SHOT`

[4] `http://graphics.stanford.edu/data/3Dscanrep`

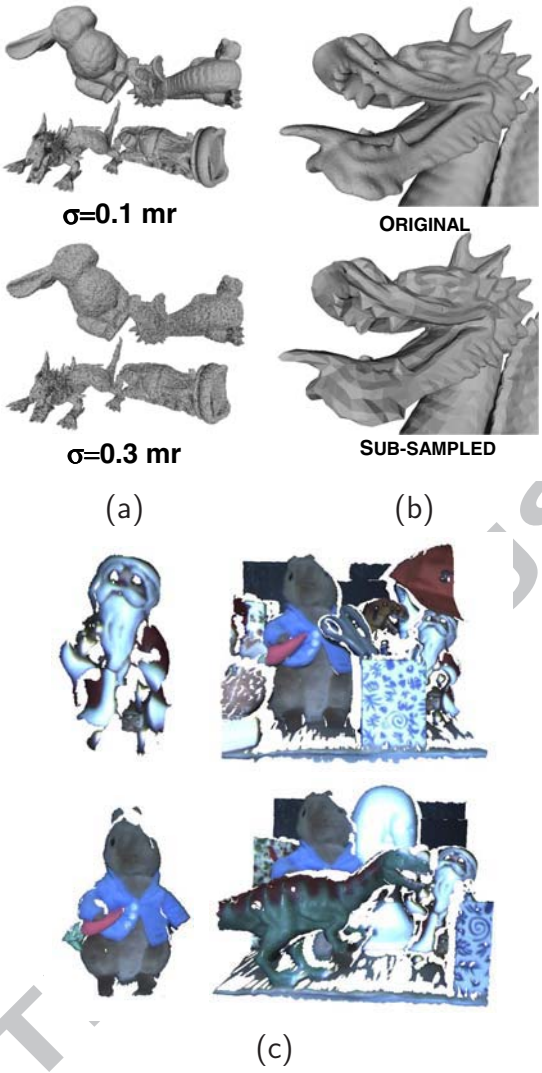


Figure 5: a) Dataset 1: one scene at the 2 noise levels; b) Dataset 2: a detail from a scene with and without sub-sampling; c) Dataset 3: two models and two scenes.

The parameter values of the considered methods are kept fixed throughout all experiments, with some of them chosen by means of a tuning process. More precisely, for each descriptor we tuned the *support radius* and those parameters (either one or two) influencing the *length of the descriptor*. We did not tune the length of the descriptor only for FPFH, because it is fixed

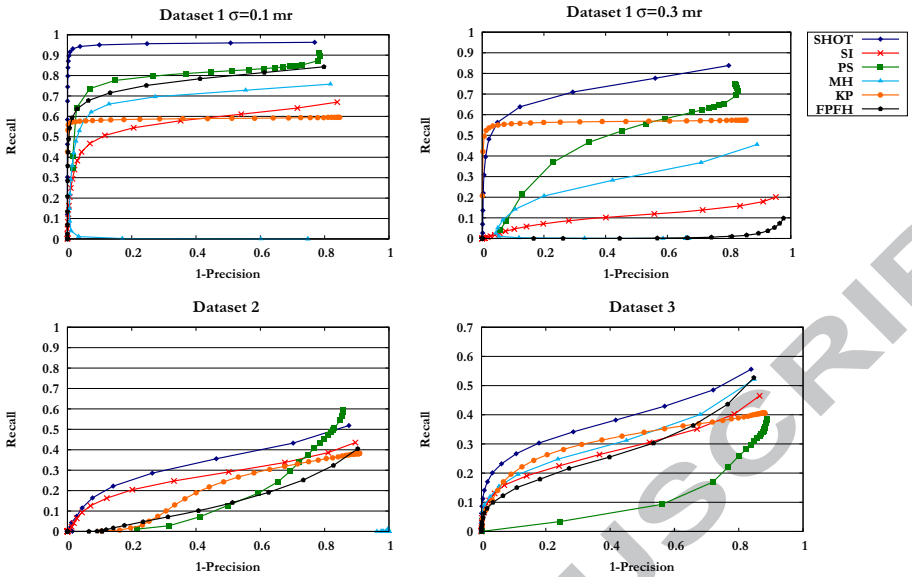


Figure 6: Precision-Recall curves regarding the *Original Local RF* experiment on Datasets 1, 2, 3.

in its PCL implementation. The tuning process consisted of a grid search within the parameter space based on a tuning scene built by rotating and translating three Stanford models ("Bunny", "Happy Buddha", "Dragon") and corrupting them with noise level σ_1. As for additional parameters, we used the default values indicated by the authors in the original paper or implementation. The parameter values used throughout the experiments are reported in Table 2.

Results for the three Datasets for the Original Local RF experiment are shown in Figure 6. SHOT notably outperforms all other methods at both noise levels on Dataset 1. It is worth observing that as for robustness to noise, KP is definitely the most effective approach, as its performance is almost unchanged at the different noise levels: we ascribe such robustness to the smoothing of the support performed within the algorithm by *gridfit*. On the other hand, such smoothing hinders the descriptiveness of the method at low noise levels (*i.e.* σ_1). This can be seen by comparing KP with the other evaluated signature PS: KP yields a worse performance in the presence of low noise (*i.e.* σ_1), but outperforms PS at noise level σ_3. As for SI and FPFH, they appear to be highly sensitive to noise, their performance notably dete-

riorating as the noise level increases. This is due to SI being highly sensitive to small variations in the normal estimation (*i.e.* SI Reference Axis), that we compute as proposed in [19]. This is consistent with the results reported by [17]. FPFH is even more sensitive to noise on normals than SI: this is reasonable, as FPFH does not use the normal at the keypoint as a reference axis only, but also uses the normals at the points in the support to compute the angles with the Reference Axis that fill the histogram. Finally, although MH shares the same hybrid structure as SHOT, it cannot successfully deal with noise due to its deployment of curvatures, which are very susceptible to noise as they require computation of second-order derivatives.

As for Dataset 2, it is clear that point density variation is a very challenging nuisance, causing a severe performance loss to all methods. SHOT and SI yield comparable performance, which turns out overall superior than that yielded by the other descriptors. Interestingly, between less effective methods there is KP, which deploys specific approaches to counteract point density variations, whose effectiveness could be however washed out by the instability of the local RF in the presence of such a nuisance.

Fig. 6 highlights that SHOT is significantly more effective than other methods also in Dataset 3. According to the previously discussed characteristics of the dataset, this demonstrates the higher descriptiveness embedded into SHOT compared to the other methods. The performance of MH, which turns out, together with KP, the second-best method in this experiment shows that a signature of histogram structure holds the potential to attain a distinctive descriptor: encoding curvatures, *i.e.* second order derivatives, on this dataset turns out effective given the lower noise level of Spacetime data. FPFH and SI are also effective because normals, alike, can be estimated more reliably. As for Signatures, our analysis indicates that KP is limited in its performance by its not unique local RF, while PS is not as effective as expected on this dataset, clearly yielding the worst performance. This seems to be due to the fact that PS tends to describe too few points (*i.e.* those laying at the intersection between the sphere around a feature and the surface) to be able to capture enough distinctive shape information in presence of smooth surfaces.

The results of the *Same local RF* experiment concerning Datasets 1, 2 and 3 are reported in Fig. 7. These results aim at highlighting robustness and descriptiveness of descriptors independently of the repeatability of the algorithm adopted to establish a canonical local reference (for those methods requiring a Reference Axis only, we use the **z** axis of our RF). As for

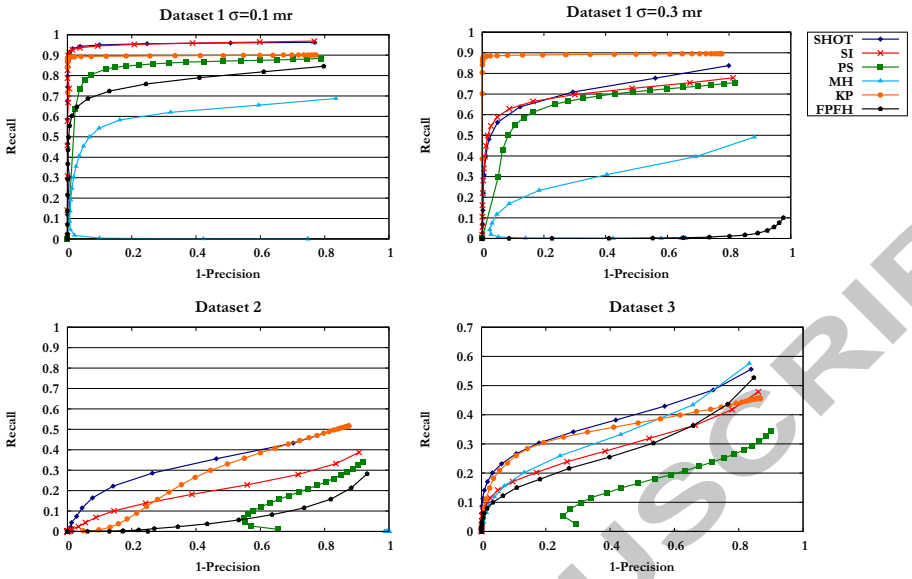


Figure 7: Precision-Recall curves regarding the *Same Local RF* experiment on Datasets 1, 2, 3.

Dataset 1, coherently with previous experiments, the most robust method is KP, which also achieves the best performance at the higher noise level (σ_3). SHOT and SI are the best methods at the lower noise level (σ_1) and the second bests at the higher noise level. These results confirm that the unsatisfactory performance of KP and SI in the previous experiment (Fig. 6) are due to, respectively, the not uniqueness of the local RF proposed in [18] and the instability of the normal estimation used in [19]. FPFH, instead, does not benefit from a more repeatable RA, its performance only slightly improving with respect to the previous experiment: unlike SI, it accumulates angles between normals and the RA, and normal directions are more affected by noise compared to the plain point coordinates deployed by SI. As far as sigantures are concerned, in the ideal case of absence of noise, they represent the most descriptive design, hence are expected to yield better performance. Indeed, PS performance improves with respect to the previous experiment: signatures benefit from a robust local RF, which approximates better the ideal, noiseless case. However, PS is still significantly less effective than SHOT and SI on this dataset: even when equipped with a robust LRF, signatures are sensitive to noise and have to deploy costly smoothing procedures (as done

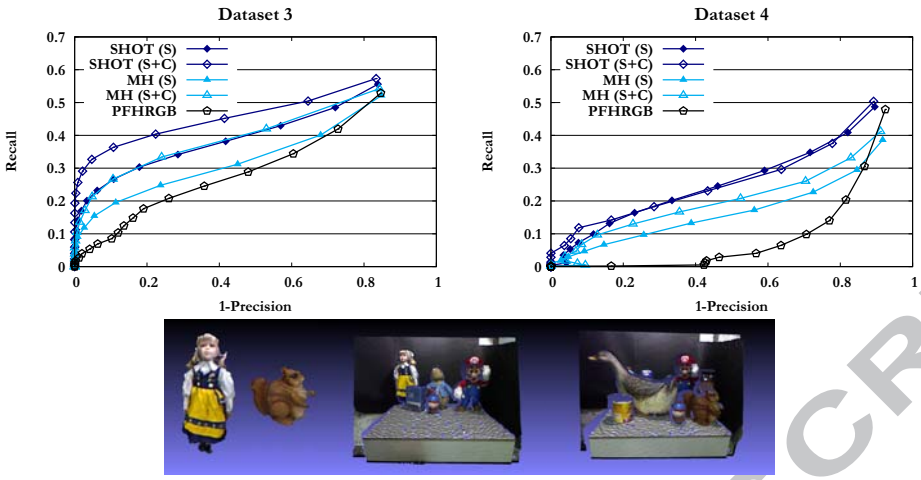


Figure 8: Top: Comparison between "color+shape" descriptors (SHOT (S+C), MH (S+C)) and their "shape-only" counterparts (SHOT (S), MH (S)) on datasets 3 and 4. Bottom: models (leftmost figure) and scenes (rightmost figures) from Dataset 4.

by KP) to be as robust as other designs. Similarly, as for MH, the experiment highlights that a descriptor based on curvatures is inherently unsuited to noisy data, even when endowed with a robust canonical reference frame.

As for Dataset 2, PS, FPFH and SI performance deteriorates with respect to the results in Fig. 6, whereas KP slightly improves. This shows that the proposed local RF is less robust to point density variations than the canonical reference adopted by PS and normal estimation based on mesh triangles. On the other hand, it provides a unique reference that is beneficial to the specific procedures dealing with point density variations embodied into KP. As for Dataset 3, coherently with previous discussion, comparison between Fig. 6 and Fig. 7 highlights how the performance of KP is hindered by the use of a not unique local RF and can be boosted by the deployment of the method proposed in Sec. 3.

5.1.2. Experiments on RGB-D data

The experimental results shown in this section are aimed at evaluating SHOT in descriptor matching tasks dealing with RGB-D data. To distinguish between the two versions of SHOT (*i.e.* that describing only Shape and that relying on both Shape and Color) we will refer to them as SHOT (S) and

SHOT (S+C), respectively. Thus, SHOT (S+C) is compared to SHOT (S) to assess the benefits brought in by the combined deployment of texture and shape. Moreover, we compare our proposal to MeshHoG and PFHRGB[5], which, to the best of our knowledge, are the only approaches other than SHOT combining both color and shape within the same description scheme. Similarly to SHOT, we consider two versions of MeshHoG: one deploying only Shape and the other relying on both Shape and Color, which will be denoted as MH (S) and MH (S+C) respectively. For MH (S) we use the mean curvature as shape cue. As described in [11], the use of both shape and color can be achieved by juxtaposing two MeshHoG descriptors relying respectively on mean curvature and color. It is worth pointing out that, unlike the findings reported in [11], with our datasets the shape-and-color MeshHoG, i.e. MH (S+C), provides slightly better results than the color-only version. Therefore, we make use of MH (S+C) in our experimental evaluation.

Since both SHOT (S+C) and MH (S+C) introduce an additional parameter related to color description, a specific tuning of these two parameters was performed on a single RGB-D scene (not included then in the test datasets) by keeping all other parameters to the values used throughout previous experiments. The tuned values of these additional parameters are also included in Table 2 (*Histogram Color Bins*, fifth column). As it was the case with FPFH, we could not tune the length of PFHRGB since it is fixed in its PCL implementation. We first compare the methods on Dataset 3 (Spacetime Stereo)[6]. Fig. 8 reports the results. The color-enhanced versions of SHOT and MH are always able to outperform their respective shape-only counterparts. This result highlights the importance of relying on the color cue when available to improve the effectiveness of 3D descriptors. SHOT (S+C) is more effective than MH (S+C), resulting overall the best descriptor in terms of descriptiveness and robustness, delivering state-of-the-art performance on the considered dataset. Similarly to shape-only dataset, PFHRGB performs worse than SHOT and MH: this may be due to the sensitivity of (F)PFH schemes to noisy normal estimation, which was highlighted by previous experiments. It is worth noting that this is not in contrast with the results provided in

[5]an extension of the PFH descriptor [27] to include RGB data in the description, available in PCL.

[6]Datasets 1 and 2 cannot be used here since they do not include color information

the evaluation performed by Alexandre [43]: in that case, PFHRGB turned out the best descriptor, followed by SHOT (S+C). Indeed, these results were obtained on a dataset targeting category recognition by object views without any clutter and occlusions, while our experiments address object detection in presence of clutter and occlusions.

We also evaluated the performance of descriptors on RGB-D data acquired with the Microsoft Kinect sensor. This novel dataset, hereinafter referred to as Dataset 4 and publicly available through the SHOT project page[7], includes several views of 6 models and 15 scenes. Examples of model views and scenes are shown in Fig. 8. The depth estimation approach of the Kinect sensor allows for higher frame rates than the Spacetime Stereo method used for Dataset 3, but data are significantly noisier, resulting in the most challenging dataset between those used in our experiments.

Indeed, although we do not report the charts here for the sake of space, we found that if descriptors are evaluated on the Kinect dataset using the parameter values adopted so far (left hand side of the / symbol in Tab. 2), all perform poorly: *e.g.* for SHOT (S+C), which still provides the best performance on Kinect data, Recall raises above 0.15 only with Precision less than 0.3. Therefore, we ran a specific tuning process for Dataset 4. The resulting parameter values are reported in Table 2 (right hand side of the / symbol). As can be noted, the length of descriptors shrinks, with usually fewer bins for each histogram, and radii get larger, a clear indication of the need to increase the coarseness of the representation to deal with noisier data.

Throughout the matching process we now extract a smaller number of features from each model than in previous experiments (*i.e.* 100 instead of 1000). This is due to the lower density of vertices per unit of volume which characterizes the data acquired by the Kinect sensor with respect to the Spacetime Stereo set-up. The lower density, in turn, is due to the different minimum acquisition distance of the two sensors: the Kinect requires objects to be at least a meter away from the device, whereas the baseline of the camera pair we used in the Spacetime Stereo setup is small enough to correctly estimate depth of nearer objects. With less vertices per unit of volume, 1000 feature points result in too dense a sampling of the model surface, which may cause many more model features to be similar than in previous experiments. Since scene features are matched to model features, a relatively high number

[7] `www.vision.deis.unibo.it/SHOT`

of similar model features is detrimental to the adopted matching criterion based on the ratio between the first and second nearest neighbor descriptor distances (see Sec. 5).

Results are reported in Fig. 8. As it can be clearly seen in the Figure, deployment of the color cue improves again the performance of both MeshHoG and SHOT. Similarly, PFHRGB delivers the worst performance: we ascribe this to the high noise level of the Kinect data, which makes it extremely difficult to obtain repeatable normal estimation. Overall, SHOT performs better than other approaches considering both color and shape also on the very challenging RGB-D data provided by the Kinect.

5.1.3. Computational efficiency

We have also compared descriptors in terms of their computational efficiency. It is important to note that certain algorithms cannot deploy efficient indexing schemes (*e.g.* kD-tree) in order to speed-up the matching stage. This is the case *e.g.* of PS, due to the non metric nature of its matching measure. Hence, we measure execution times using, for all methods, a *Brute Force* matching algorithm. Results are shown in Fig. 9: the two charts report the number of milliseconds per correspondence needed by the various methods using different support sizes. In addition, they also highlight, by a red dot along each curve, the working point specific to each algorithm referred to the parameter set obtained by the tuning process described in Sec. 5.1.1.

Overall, these results demonstrate the notable differences in computational efficiency between the algorithms. In particular, at the tuned support sizes, SI, PS and SHOT run one or more orders of magnitude faster than KP,MH and FPFH, SI turning out consistently slightly faster than SHOT at each support size. With regards to PS, the use of multiple local RFs mainly accounts for slowing down the matching stage. As previously mentioned, our implementation of KP runs partly in MATLAB, in particular the *gridfit* script. However, this does not invalidate the comparison because this hybrid version is the fastest implementation available. In fact, we tried to port the *gridfit* script to C++, in order to have commensurable implementations, and it turned out that the version using the MATLAB script is one order of magnitude faster, due to its highly optimized linear algebra libraries, which are impossible to replicate in a custom implementation. As for shape and color descriptors, it can be noted that SHOT (S+C) is approximately three times slower than SHOT (S) , one order of magnitude faster than MH (S+C) and

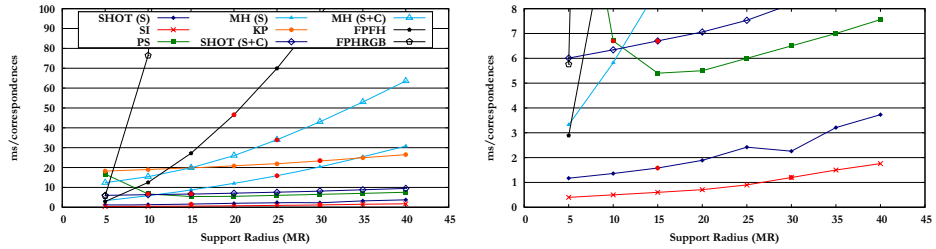


Figure 9: ms/correspondence vs. support radius (in the right chart the time axis is zoomed in for better comparison between SI and SHOT).

two orders of magnitude faster than PFHRGB.

5.2. Retrieval experiment

The second experiment compares SHOT to KP, MeshHoG, PS, SI and FPFH in a typical shape retrieval scenario based on the popular *Princeton Shape Benchmark* dataset [44] at its *coarse1* classification level, which contains 1814 "depth-only" models, split into a train set and a test set. Of course, shape retrieval may be solved with many different approaches, even more effective than the use of local features, due to the absence of clutter and noise (see [45] for a detailed survey): the aim of this experiment is only to evaluate the relative performance of the considered local features assuming that a local feature approach to shape retrieval has been selected. Specifically, for each query we establish a set of correspondences by matching 200 descriptors at randomly extracted positions to their 1-NNs within the descriptors extracted at the same number of random positions on each training model. Models are ranked according to the mean distance of their descriptors from the query. Moreover, we apply a common pre-processing step which includes normalizing the size of all shapes to the unit sphere centered in the origin, aligning them to their principal directions and re-sampling to a pre-defined number of vertices (4000). As for parameters, all descriptors are evaluated using three different radius values, *i.e.* 0.1, 0.3 and 0.5, while the remaining parameters are set to the same values as in the descriptor matching experiment (see Table 2). We employ the *normalized Discounted Cumulative Gain (nDCG)* [44] as performance index. DCG weights correct results near the front of the list more than correct results later in the ranked list. To obtain the nDCG, the DCG is normalized with respect to the average over all algorithms tested and the average is shifted to zero: hence,

positive/negative nDCG scores represent above/below average performance, and higher numbers are better. The results in Fig. 11 show that SHOT and FPFH yield the best results, SHOT turning out superior when using two of the three considered radii and achieving the highest performance among all the descriptors on this dataset (at radius 0.5).

5.3. Reconstruction experiments

This last group of qualitative experiments addresses a 3D reconstruction scenario, where we demonstrate the versatiblity of SHOT by using it also to perform fully-automatic 3D reconstruction from unordered 2.5D views. We process data gathered by two different sensors: a Spacetime Stereo system and a Microsoft Kinect. Although recent approaches can register in real-time successive views from a Kinect sensor [46] by exploiting temporal coherency and thus performing SLAM, to the best of our knowledge, fully automatic 3D reconstruction from unordered Spacetime Stereo and Kinect views has not been demonstrated yet.

In both experiments, fully automatic 3D reconstruction is achieved by a two-step procedure: first, a coarse registration is obtained by matching SHOT descriptors, estimating the 3D Euclidean transformations between every pair of views and retaining only those view pairs maximizing the global area of overlap; this coarse registration is then fed as initial guess to a final ICP registration. Maximization of the area of overlap is achieved through the Maximum Spanning Tree approach described in [8].

In the first experiment, two objects are reconstructed, each from a set of 2.5 Spacetime views covering a full 360° field of view around the object. The first two rows of Fig. 10 show the reconstructions obtained for the two objects by relying on depth measurements only, *i.e.*. by means of SHOT (S). As vouched by these qualitative results, without any assumptions on the initial poses, SHOT correspondences allows for achieving a coarse alignment which results in an initial guess accurate enough to successfully reconstruct the 3D shape of the object without any manual intervention.

In the second experiment, several RGB-D views around two objects have been acquired with a Kinect. The last two rows of Fig. 10 provide qualitative results concerning the two reconstructions, obtained by describing both shape and texture, *i.e.*. by means of SHOT (S+C). We can observe that also with the very noisy RGB-D data yielded by the Kinect sensor, SHOT allows for accurate fully-automatic 3D reconstruction from unordered views.

Figure 10: 3D Reconstruction from Spacetime Stereo views (first two rows) and Kinect views (last two rows): (a) initial set of views (b) coarse registration (c) global registration frontal view (d) global registration additional view.

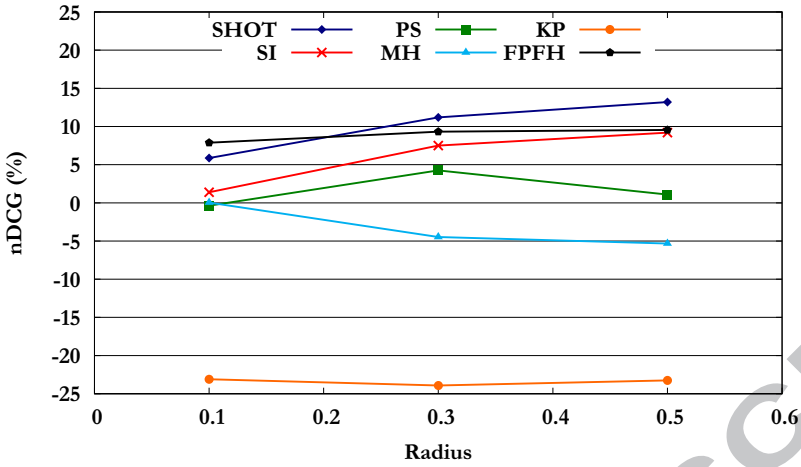


Figure 11: Results for the shape retrieval experiment.

5.4. Applications

The experiments proposed in this papers are focused on comparing SHOT to state-of-the art 3D descriptors in either plain matching tasks or within baseline pipelines, so as to achieve independence from specific additional stages that may mask the performance of evaluated descriptors. Nevertheless, it is worth pointing out that recently SHOT has been successfully employed within state-of-the-art 3D pipelines aimed at 3D object recognition in clutter and occlusion. This is the case of [47], where the authors report their choice of SHOT as 3D descriptor because of the superior performance with respect to the other considered approaches, as well as of [48], whose proposed SHOT-based 3D pipeline was able to solve the benchmark dataset introduced in [18].

Moreover, SHOT has been included in several recent evaluations of 3D descriptors aimed at different application scenarios, providing in most cases notable performance. As for object recognition in clutter from RGB-D data, SHOT (S+C) turned out the second-best approach in terms of recognition capability in the evaluation proposed in [43], being also much more efficient than the PFHRGB, the top performer. In the evaluation proposed in [49], focused on object recognition in clutter and occlusion from CAD models, SHOT (S) is the second-best performer, turning out slightly less accurate than the top method, which is however hindered by a huge memory footprint. In the comparison presented in Proença et al. [50], SHOT outperforms Spin

Images and dense SIFT for the task of object category recognition on the RGB-D Object dataset [51]. Likewise, Behley et al. [52] show that SHOT yields good performance in the context of classification of urban data acquired from LIDARs, resulting overall one of the two best performers. SHOT has also been used in the context of medical image analysis, proving to be one of the best descriptors - together with 3DSC and SI - for the task of automatic localization of 3D craniofacial landmarks [53]. As for 3D registration aimed at visual verification of correct endpoint alignment for the humanoid Justin [54], SHOT provided the best registration accuracy. Finally, SHOT has been deployed also as a global descriptor in the context of 3D object retrieval from a large 3D CAD database [55]: in the proposed comparison to other global descriptors it turned out the second-best approach in terms of retrieval accuracy.

6. Concluding remarks

The SHOT descriptor provides an effective and efficient tool for surface matching applications. It is rotation invariant and robust to noise as well as spurious shape variations due to low quality sensors. Such robustness does not affect distinctiveness though, SHOT featuring the ability to capture the important traits of the underlying shape. The descriptor has been designed to lay at the intersection between the two classes gathering the vast majority of works in the area of 3D description, namely Signatures and Histograms. A thorough experimental evaluation demonstrates that SHOT compares favorably to state-of-the-art 3D description approaches. Moreover, experimental results validate the proposed categorization as well as the intuition that the synergy between the design of a unique and repeatable local RF and the embedding of an hybrid signature/histogram structure into a descriptor allows for achieving at the same time state-of-the-art robustness and descriptiveness. The SHOT design allows for seamlessly integrating description of texture to improve distinctiveness. This qualifies SHOT as a natural tool to carry out surface matching based on the output of the increasingly available low-cost RGB-D sensors, such as the already widespread Microsoft Kinect.

SHOT can be easily used and evaluated by other researchers, as also discussed in sec. 5.4, for its implementation has been included into the open-source *Point Cloud Library* (PCL) and is also available through the official

project page[8].

Several interesting directions for further research stem from this work. Point density variation proves to be a major cause of failure for 3D descriptors. However, it occurs quite often with real data, due to the decreasing depth resolution at larger distances from the sensor. Another issue, which is peculiar to the 3D scenario with respect to the field of 2D features, concerns the handling of missing parts within the input mesh. Real 3D data are likely to have holes and irregular borders. At present, no specific mechanism has been proposed in literature to deal with this issue in the description stage. In this respect, SHOT's structure may allow for handling missing data during the matching stage by deploying a suitable metric to compare descriptors, such as the Earth Mover's Distance (EMD) [56] or the recently proposed Quadratic-Chi Histogram Distance [57]. Another relevant research line concerns real time computation of 3D descriptors. Although SHOT and Spin Images are remarkably fast, the wide-spreading of real-time 3D sensors calls for faster than real-time feature extraction and description, so as to allow enough time for subsequent processing, *e.g.* for object recognition or SLAM. In this respect, more efficient algorithm to compute SHOT descriptors may be devised when the input surface is provided as a range image, *i.e.* as a regular lattice of X, Y and Z coordinates, or when the input mesh is voxelized.

References

1. Zhao, W., Chellappa, R., Phillips, P., Rosenfeld, A.. Face recognition: A literature survey. *ACM Computing Survey* 2003;**35**(4):399–458.

2. Iyer, M., Jayanti, S., Lou, K., Kalyanaraman, Y., Ramani, K.. Three dimensional shape searching: State-of-the-art review and future trends. *Computer Aided Design (CAD)* 2005;**5**(15):509–530.

3. Ovsjanikov, M., Sun, J., Guibas, L.. Global intrinsic symmetries of shapes. *Computer Graphics Forum* 2008;**5**:1341–1348.

4. Somanath, G., Kambhamettu, C.. Abstraction and generalization of 3D structure. In: *Proc. of the Asian Conference on Computer Vision (ACCV) - Part III*; Lecture Notes in Computer Science. Springer-Verlag, Berlin, Heidelberg; 2011, p. 483–496.

[8] `www.vision.deis.unibo.it/SHOT`

5. Mian, A.S., Bennamoun, M., Owens, R.A.. On the repeatability and quality of keypoints for local feature-based 3D object retrieval from cluttered scenes. *International Journal of Computer Vision (IJCV)* 2010; **89**(2-3):348–361.

6. Chen, H., Bhanu, B.. 3D free-form object recognition in range images using local surface patches. *Pattern Recognition Letters* 2007; **28**(10):1252–1262.

7. Zhong, Y.. Intrinsic shape signatures: A shape descriptor for 3D object recognition. In: *Proc. of the Int. Conf. on Computer Vision (ICCV) - 3D Representation for Recognition Workshop (3dRR)*. IEEE Computer Society Washington, DC, USA; 2009, p. 689–696.

8. Novatnack, J., Nishino, K.. Scale-dependent/invariant local 3D shape descriptors for fully automatic registration of multiple sets of range images. In: *Proc. of the European Conference on Computer Vision (ECCV)*. Springer-Verlag, Berlin, Heidelberg; 2008, p. 440–453.

9. Unnikrishnan, R., Hebert, M.. Multi-scale interest regions from unorganized point clouds. In: *Proc. of the Int. Conf. on Computer Vision and Pattern Recognition (CVPR) - Workshop on Search in 3D (S3D)*. IEEE Computer Society Washington, DC, USA; 2008, .

10. Akagunduz, E., Ulusoy, I.. 3D object representation using transform and scale invariant 3D features. In: *Proc. of the Int. Conf. on Computer Vision (ICCV)*. IEEE Computer Society Washington, DC, USA; 2007, p. 1–8.

11. Zaharescu, A., Boyer, E., Horaud, R.. Keypoints and local descriptors of scalar functions on 2d manifolds. *International Journal of Computer Vision (IJCV), to appear* 2012;.

12. Tombari, F., Salti, S., Di Stefano, L.. Performance evaluation of 3d keypoint detectors. *International Journal of Computer Vision (IJCV), to appear* 2012;.

13. Tombari, F., Salti, S., Di Stefano, L.. Unique signatures of histograms for local surface description. In: *Proc. of the 11th European Conference on Computer Vision (ECCV)*. 2010, .

14. Petrelli, A., Di Stefano, L.. On the repeatability of the local reference frame for partial shape matching. In: *Proc. of the Int. Conf. on Computer Vision (ICCV)*. 2011, p. 2244–2251. doi:\bibinfo{doi}{10.1109/ICCV.2011.6126503}.

15. Tombari, F., Salti, S., Di Stefano, L.. A combined intensity-shape descriptor for texture-enhanced 3D feature matching. In: *Proc. of the 18th Int. Conf. on Image Processing (ICIP)*. 2011, .

16. Rusu, R., Blodow, N., Beetz, M.. Fast point feature histograms (fpfh) for 3d registration. In: *Proc. of the Int. Conf. on Robotics and Automation (ICRA)*. 2009, .

17. Frome, A., Huber, D., Kolluri, R., Bülow, T., Malik, J.. Recognizing objects in range data using regional point descriptors. In: *Proc. of the European Conference on Computer Vision (ECCV)*; vol. 3. 2004, p. 224–237.

18. Mian, A., Bennamoun, M., Owens, R.. A novel representation and feature matching algorithm for automatic pairwise registration of range images. *International Journal of Computer Vision (IJCV)* 2006;**66**(1):19–40.

19. Johnson, A., Hebert, M.. Using spin images for efficient object recognition in cluttered 3D scenes. *Trans on Pattern Analysis and Machine Intelligence (PAMI)* 1999;**21**(5):433–449.

20. Stein, F., Medioni, G.. Structural indexing: Efficient 3-d object recognition. *Trans on Pattern Analysis and Machine Intelligence (PAMI)* 1992;**14**(2):125–145.

21. Chua, C.S., Jarvis, R.. Point signatures: A new representation for 3D object recognition. *International Journal of Computer Vision (IJCV)* 1997;**25**(1):63–85.

22. Sun, Y., Abidi, M.A.. Surface matching by 3D point's fingerprint. *Int Conf on Computer Vision (ICCV)* 2001;**2**:263–269.

23. Darom, T., Keller, Y.. Scale-invariant features for 3-D mesh models. *Image Processing, IEEE Transactions on* 2012;**21**(5):2758–2769. doi:\bibinfo{doi}{10.1109/TIP.2012.2183142}.

24. Knopp, J., Prasad, M., Willems, G., Timofte, R., Van Gool, L.. Hough transform and 3D SURF for robust three dimensional classification. In: *ECCV*. 2010, .

25. Bay, H., Ess, A., Tuytelaars, T., Van Gool, L.J.. Speeded-Up Robust Features (SURF). *Computer Vision and Image Understanding* 2008; **110**(3):346–359.

26. Dorai, C., Jain, A.. COSMOS-a representation scheme for 3D free-form objects. *Trans on Pattern Analysis and Machine Intelligence (PAMI)* 1997;**19**(10):1115–1130.

27. Rusu, R., Blodow, N., Marton, Z., Beetz, M.. Aligning point cloud views using persistent feature histograms. In: *Proc. of the Int. Conf. on Intelligent Robots and Systems (IROS)*. 2008, .

28. Guo, Y., Sohel, F., Bennamoun, M., Lu, M., Wan, J.. Rotational projection statistics for 3d local surface description and object recognition. *International Journal of Computer Vision* 2013;**105**(1):63–86.

29. Sun, J., Ovsjanikov, M., Guibas, L.. A concise and provably informative multi-scale signature based on heat diffusion. In: *Proc. Symp. Geom. Proc.* 2009, p. 1383–1392.

30. Knopp, J., Prasad, M., Van Gool, L.. Orientation invariant 3D object classification using hough transform based methods. In: *ACM Multimedia 2010 Workshop - 3D Object Retrieval*. 2010, .

31. Leibe, B., Leonardis, A., Schiele, B.. Robust object detection with interleaved categorization and segmentation. *International Journal of Computer Vision (IJCV)* 2008;:17–32.

32. Hoppe, H., DeRose, T., Duchamp, T., McDonald, J., Stuetzle, W.. Surface reconstruction from unorganized points. In: *Proc. of the Conference on Computer Graphics and Interactive Techniques (SIGGRAPH)*. ACM, New York, NY, USA; 1992, p. 71–78.

33. Mitra, N.J., Nguyen, A., Guibas, L.. Estimating surface normals in noisy point cloud data. *International Journal of Computational Geometry and Applications* 2004;**14**(4–5):261–276.

34. Bro, R., Acar, E., Kolda, T.. Resolving the sign ambiguity in the singular value decomposition. *Journal of Chemometrics* 2008;**22**:135–140.

35. Lowe, D.G.. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision (IJCV)* 2004;**60**:91–110.

36. Fairchild, M.. *Color Appearance Models*. John Wiley & Sons Ltd., Chichester, UK; 2005.

37. Mikolajczyk, K., Schmid, C.. A performance evaluation of local descriptors. *Trans on Pattern Analysis and Machine Intelligence (PAMI)* 2005;**27**(10):1615–1630.

38. Ke, Y., Sukthankar, R.. PCA-SIFT: A more distinctive representation for local image descriptors. In: *Proc. of the Int. Conf. on Computer Vision and Pattern Recognition (CVPR) - Volume 2*. IEEE Computer Society, Washington, DC, USA; 2004, p. 506–513.

39. D'Errico, J.. Surface fitting using gridfit. MATLAB Central File Exchange; 2010.

40. Conde, C., Rodrguez-Aragn, L., Cabello, E.. Automatic 3D face feature points extraction with spin images. *Int Conf on Image Analysis and Recognition (ICIAR)* 2006;**4142**:317–328.

41. Davis, J., Nehab, D., Ramamoothi, R., Rusinkiewicz, S.. Spacetime stereo: A unifying framework for depth from triangulation. *Trans on Pattern Analysis and Machine Intelligence (PAMI)* 2005;**27**(2):1615–1630.

42. Zhang, L., Curless, B., Seitz, S.. Spacetime stereo: Shape recovery for dynamic scenes. In: *Proc. of the Int. Conf. on Computer Vision and Pattern Recognition (CVPR) - Volume 2*. IEEE Computer Society Washington, DC, USA; 2003, p. 367–374.

43. Alexandre, L.. 3d descriptors for object and category recognition: a comparative evaluation. In: *IROS Workshop on Color-Depth Camera Fusion in Robotics*. 2012, .

44. Shilane, P., Min, P., Kazhdan, M., Funkhouser, T.. The princeton shape benchmark. In: *Shape Modeling International*. 2004, .

45. Tangelder, J.W., Veltkamp, R.C.. A survey of content based 3d shape retrieval methods. *Multimedia Tools Appl* 2008;**39**(3):441–471. doi:\bibinfo{doi}{10.1007/s11042-007-0181-0}. URL `http://dx.doi.org/10.1007/s11042-007-0181-0`.

46. Izadi, S., Kim, D., Hilliges, O., Molyneaux, D., Newcombe, R., Kohli, P., et al. Kinectfusion: real-time 3D reconstruction and interaction using a moving depth camera. In: *Proc. of the 24th annual ACM symposium on User interface software and technology (UIST)*. 2011, p. 559–568.

47. Rodolá, E., Albarelli, A., Bergamasco, F., Torsello, A.. A scale independent selection process for 3d object recognition in cluttered scenes. *International Journal of Computer Vision (IJCV), to appear* 2012;.

48. Aldoma, A., Tombari, F., Di Stefano, L., Vincze, M.. A global hypotheses verification method for 3d object recognition. In: *ECCV*. 2012, .

49. Aldoma, A., Marton, Z., Tombari, F., Wohlkinger, W., Potthast, C., Zeisl, B., et al. Point cloud library: Three-dimensional object recognition and 6 dof pose estimation. *IEEE Robotics and Automation Magazine (RAM)* 2012;**19**(3):80–91.

50. Proença, P.F., Gaspar, F., Dias, M.S.. Good appearance and shape descriptors for object category recognition. In: *Advances in Visual Computing*; vol. 8033 of *Lecture Notes in Computer Science*. Springer Berlin Heidelberg; 2013, p. 385–394.

51. Lai, K., Bo, L., Ren, X., Fox, D.. A large-scale hierarchical multi-view RGB-D object dataset. In: *ICRA*. 2011, p. 1817–1824.

52. Behley, J., Steinhage, V., Cremers, A.. Performance of histogram descriptors for the classification of 3d laser range data in urban environments. In: *Int. Conf. on Robotics and Automation (ICRA)*. 2012, .

53. Sukno, F., Waddington, J., Whelan, P.. Comparing 3d descriptors for local search of craniofacial landmarks. In: *Int. Symp. on Visual Computing (ISVC)*. 2012, .

54. Figueroa, N., Ali, H., Schmidt, F.. 3d registration for verification of humanoid justins upper body kinematics. In: *Int. Conf. on Computer and Robot Vision (CRV)*. 2012, .

55. Wohlkinger, W., Aldoma, A., Rusu, R., Vincze, M.. 3dnet: Large-scale object class recognition from cad models. In: *Int. Conf. on Robotics and Automation (ICRA)*. 2012, .

56. Rubner, Y., Tomasi, C., Guibas, L.J.. A metric for distributions with applications to image databases. In: *Proc. of the Int. Conf. on Computer Vision (ICCV)*. 1998, .

57. Pele, O., Werman, M.. The quadratic-chi histogram distance family. In: *Proc. of the European Conference on Computer Vision (ECCV)*. 2010, .

- This paper presents a local 3D descriptor for surface matching dubbed SHOT.
- Our proposal includes a repeatable local reference frame as well as a 3D descriptor
- It enables seamless description of shape and color data from RGB-D sensors
- It is validated in object recognition, 3D reconstruction and shape retrieval scenarios
- SHOT offers superior effectiveness with remarkably good computational efficiency